### 1 Topic

# Studying the influence of two approaches for the treatment of missing values (univariate imputation and listwise deletion) on the performances of the logistic regression.

The handling of missing data is a difficult problem. Not because of its management which is simple, we just report the missing value with a specific code, but rather because of the consequences of their treatment on the characteristics of the models learned on the treated data.

We have already analyzed this problem in a previous paper. We studied the impact of the different techniques of missing values treatment on a decision tree learning algorithm (C4.5; Quinlan 1993)<sup>1</sup>. In this paper, we repeat the analysis by examining their influence on the results of the logistic regression. In a first time, we mainly use the **R** software, with the **glm()** procedure. In a second time, we examine the behavior of the tools proposed by **Orange, Knime** and **RapidMiner**.

The origin of the missing value<sup>2</sup> is the first characteristic which determines their treatment. It can be **MCAR** (missing completely at random) i.e. the apparition of the missing value does not depend neither to the values of the variable, nor to the values of the other variables of the database. It can be **MAR** (missing at random) i.e. its occurrence does not depend on the values of the variable after we remove the influence of the other variables. For instance, managers are less likely to declare their salaries in a survey. But in this category, the probability of the occurrence of missing value does not depend on the value of the salary. Last, it can be **MNAR** (missing not at random) i.e. the occurrence of the missing value is related to the true value of the variable. We cannot use a statistical approach for the imputation in this case.

Among these configurations, the MCAR case is the easier to manage. We get unbiased estimates of parameters, even when we use very simplistic techniques such as the listwise deletion  $approach^3$ .

The characteristic of the subsequent statistical method applied on the pretreated dataset is the second important piece of the puzzle. Clearly, the influence of missing values and their treatment is not the same on a predictive analysis (e.g. logistic regression), on a clustering algorithm (e.g. an Agglomerative Hierarchical Clustering) or on a factor analysis (e.g. a principal component analysis). The distinction is all the more important that some statistical methods incorporate natively a strategy for the handling of missing values (e.g. the NIPALS algorithm<sup>4</sup> for principal components analysis and the PLS regression).

Last, the criteria for assessing the treatment of missing data is the final piece of the puzzle. The goal is not to find the "true" values of missing observations for each variable. But rather to propose replacement values (in the case of imputation) which do not alter the results of the study we are conducting. We highlight often bias and the variance of the estimated parameters in the statistical

<sup>&</sup>lt;sup>1</sup> http://data-mining-tutorials.blogspot.fr/2009/11/handling-missing-values-in-sipina.html

<sup>&</sup>lt;sup>2</sup> <u>http://www.uvm.edu/~dhowell/StatPages/More\_Stuff/Missing\_Data/Missing.html</u>

<sup>&</sup>lt;sup>3</sup> <u>http://en.wikipedia.org/wiki/Listwise\_deletion</u>

<sup>&</sup>lt;sup>4</sup> <u>http://en.wikipedia.org/wiki/Non-linear\_iterative\_partial\_least\_squares</u>

literature. But in the specific context of predictive analysis, we can also ask the question of performance. What is the missing value processing technique which enables to build the most efficient model in prediction?

In this tutorial, we consider the following configuration: (1) missing values are MCAR, we wrote a program which removes randomly some values in the learning sample; (2) we apply logistic regression on the pre-treated training data i.e. on a dataset on which we apply a missing value processing technique; (3) we evaluate the different techniques of treatment of missing data by observing the accuracy rate of the classifier on a separate test sample which has no missing values.

In a first time, we conduct the experiments with R. We compare the listwise deletion approach to the univariate imputation (the mean for the quantitative variables, the mode for the categorical ones). We will see that this latter is a very viable approach in MCAR situation. In a second time, we will study the available tools in Orange, Knime and RapidMiner. We will observe that despite their sophistication, they are not better than the univariate imputation in our context.

### 2 Dataset

**Original dataset**. We use the GERMAN CREDIT DATA in our experiment<sup>5</sup>. We randomly split the data file into two sheets in an Excel workbook (**credit-german-md-simulation.xls**): CREDIT-GERMAN-TRAIN-FULL (300 instances) is used for the learning model phase; CREDIT-GERMAN-TEST (700 instances) for the test phase. For this second sample, we created once and for all the data file CREDIT-GERMAN-TEST.TXT (text file format). There are no missing values in these two samples.

0	1-1-	(~ - → =	cre	dit-ge	rman-md-s	imulation.xls [M	lode de compat	ibilité] - N	licros	oft Excel			X
Ci	Accueil	Insertion	Mise en p	age	Formule	s Données	Révision	Affichag	e	Développeur	Complémen	ts 🔞 – 🖻	۶ x
Visual Macros Basic					Insérer	Mode réation	iétés liser le code iter la boîte de d	lialogue	Sou	rce Propriét	és du mappage <b>dension</b> er les données	Exporter	
		Code				Contro	oles				XML		
	L27	<del>•</del> (*	$f_{x}$	car									×
4	D	E	F		G	Н		J		K	L	M	F
1	purpose	credit_amoun	savings_s	tatu e	mploymen	t installment_c	personal_stat	other_pa	rties	residence_sir	property_mag	age	ot
2	radio/tv	5943	no known s	avin <	1	1	female div/dep/r	none		1	car	44	i no
3	furniture/equipm	3650	<100	<	1	1	female div/dep/r	none		4	car	22	t no
4	furniture/equipm	2069	no known s	avin 1	<=X<4	2	male mar/wid	none		1	car	26	i no
5	business	1568	100<=X<50	0 1	<=X<4	3	female div/dep/r	none		4	life insurance	24	no
6	radio/tv	1533	<100	<	1	4	female div/dep/r	none		3	car	38	i sti
7	business	1941	>=1000	1	<=X<4	4	male single	none		2	life insurance	35	i no
8	radio/tv	1364	<100	4	<=X<7	3	male single	none		4	real estate	59	l no
9	used car	2993	<100	1	<=X<4	3	male single	none		2	real estate	28	i sti
10	new car	5045	no known s	avin >	=7	1	female div/dep/r	none		4	car	59	l no 🔻
• •	Credi	t-german-md	🧹 credit-	germa	an-train-full	credit-gerr	nan-test 🦯 🐮	I 4					► I
Prêt	: 🛅			1		1					100 % 🕞 —		÷:

**Creating a training sample with missing values.** We use a VBA program to generate a learning sample with some missing values. For this, we launch the program by clicking on the DEVELOPPEUR / MACROS into Excel (*for the French version of Excel*). Into the dialog box which appears, we select the STARTDATAGENERATOR program.

<sup>&</sup>lt;sup>5</sup> http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

C		€ • ) ÷	credit		simulat	ion.xls [Mode d	le compatibilité] - Micro	soft Excel		
	Accueil	Insertion	Mise en page	Formul	les	Macro		Castlene a	? ×	]- □ ×
Vi Bi	sual Macros	Enregistrer un Utiliser les réf Sécurité des m Code	e macro érences relatives iacros	insérer	Mode Créatior	Nom de la macr StartDataGene StartDataGene	o : erator erator		E <u>x</u> écuter Pas à pas détaille	orter
1 2 3	Afficher le Affiche la exécuter, radic furn Appur	es macros (Alt+) a liste des macro , créer ou supp yez sur F1 pour	F8) os que vous pou rimer. obtenir de l'aid	ivez i e.	nt inst				Modifier       Créer       Supprimer	• • • • • • • • • • • • • • • • • • •
4	furniture/equipm	2069	no known savin	1<=X<4				-	Options	26 no
5	business	1568	100<=X<500	1<=X<4						24 no
7	radio/tv	1533	<100	<1 12-X24		Macros dans :	Tous les classeurs ouverts	s 💌		35 no
8	radio/tv	1364	<100	4<=X<7		Description				59 no
9	used car	2993	<100	1<=X<4						28 st
10	new car	5045	no known savin	>=7						59 no 🔻
l∢ Prê	t ► ► credit	t-german-md	<pre>/ credit-gern</pre>	nan-train-fu	l _ a				Annuler	→ I + .,:

The program is started, a settings dialog asks the proportion of the missing values that we want to insert into the learning sample. For instance, if we set 0.05 (5%), the program generates a learning sample with (0.05 \* 300 rows \* 20 variables = 300 cells) missing values (the '?' character is used for materializing the missing value). The class attribute is not concerned by this process.

Aissing values generator	x
Missing values proportion :	
.05	
START Cancel	

The resulting dataset is inserted in the first sheet of the workbook (CREDIT-GERMAN-MD).

0	a) 🖬 "7 -	(H ) =	crea	dit-german-md.	txt <b>de c</b>	ompatibilité] -	Microsoft Exce	1		
	Accueil	Insertion	Mise en page	Formules	Données	Révision Afr	fichage Dé	veloppeur C	ompléments	0 – 🗖 X
Pre	Coller	Bitstream Ver		E = = = = = = = = = = = = = = = = = = =	Standard	Style	™ Insérer ▼ Supprimer ▼ Format ▼ Cellules	Σ · A · Z · Trier e · filtrer Éc	t Rechercher et sélectionner • dition	
	C3	<del>,</del> (9	f <sub>x</sub> ?							*
	Α	В	С	D	E	F	G	Н		J =
1	checking_stat	duration	credit_history	, purpose	credit_amoun	savings_statu	employment	installment_c	personal_stat	other_partie
2	<0	12	critical/other ex	i new car	3499	<100	1<=X<4	3	female div/dep/r	co applicant
3	<0	6	?	furniture/equipm	1872	<100	?	4	male single	none
4	<0	6	critical/other ex	i new car	1361	<100	<1	2	male single	none
5	no checking	15	existing paid	used car	3029	<100	4<=X<7	2	male single	none
6	>=200	18	all paid	radio/tv	1445	no known savin	4<=X<7	4	male single	none
7	no checking	18	all paid	new car	6458	<100	>=7	2	male single	none
8	no checking	48	critical/other ex	i business	7629	no known savin	>=7	4	male div/sep	none
9	0<=X<200	30	critical/other ex	i new car	4249	<100	unemployed	4	male mar/wid	none
10	no checking	18	delayed previou	ubusiness	2169	<100	1<=X<4	4	male mar/wid	none
11	no checking	18	critical/other ex	i used car	?	no known savin	unemployed	2	male single	none
12	no checking	?	all paid	used car	7485	no known savin	unemployed	4	female div/dep/r	none
13	>=200	10	?	radio/tv	1347	no known savin	4<=X<7	4	?	none
14	no checking	12	critical/other ex	i new car	926	<100	unemployed	1	female div/dep/r	none
15	no checking	24	existing paid	new car	1249	<100	<1	4	male mar/wid	none
16	0<=X<200	13	existing paid	radio/tv	2101	<100	<1	2	female div/dep/r	guarantor
1/	<0	18	critical/other ex	new car	5302	<100	>=/	2	male single	none
		t-german-md	credit-gen	man-train-full	Credit-gern	nan-test 🔬 🕻				
Prê	t 🛅							100 %		

At the same time, the learning sample 'CREDIT-GERMAN-MD.TXT' (text file format) is automatically created in our working directory (the directory of the XLS file).

🔾 🔍 🗸 🦇 missing_data 🕨 md_experimen	ts_with_orange ►	german	<b>▼ 4</b> <del>9</del> <i>R</i> €	echercher dans : germar	
Organiser 👻 Inclure dans la bibliothèque 💌	Partager avec	<ul> <li>Diaporama Graver Nouveau</li> </ul>	dossier	:	- 🔟 🔞
☆ Favoris	·	Nom	Modifié le	Туре	Taille
🧮 Bureau		🚀 credit_imputation.r	29/11/2011 11:06	Tinn-R	3 Ko
Emplacements récents		credit-german-md.ows	30/11/2011 07:07	Fichier OWS	21 Ko
🝌 Téléchargements		📄 credit-german-md.txt	30/11/2011 22:46	Document texte	40 Ko
		credit-german-md-0_00.txt	29/11/2011 10:36	Document texte	41 Ko
📃 Bureau	=	📄 credit-german-md-0_50.txt	29/11/2011 10:37	Document texte	41 Ko
🥽 Bibliothèques		📄 credit-german-md-1_00.txt	29/11/2011 10:38	Document texte	41 Ko
Documents		📄 credit-german-md-2_00.txt	29/11/2011 10:38	Document texte	40 Ko
📔 Images		📄 credit-german-md-5_00.txt	29/11/2011 10:39	Document texte	40 Ko
👌 Musique		📄 credit-german-md-10_00.txt	29/11/2011 10:39	Document texte	38 Ko
📑 Vidéos		📄 credit-german-md-20_00.txt	29/11/2011 11:06	Document texte	36 Ko
🤣 Groupe résidentiel		🗐 credit-german-md-simulation.xls	03/11/2011 15:14	Feuille Microsoft	357 Ko
🥦 Maison		📄 credit-german-test.txt	03/11/2011 15:14	Document texte	95 Ko
🖳 Ordinateur		尾 process - rapidminer.png	30/11/2011 19:39	Image PNG	53 Ko
🗣 Réseau		results_experiments.xls	30/11/2011 18:58	Feuille Microsoft	9 Ko
ອ Panneau de configuration	-				

We use the following VBA code for this task.

```
Private Sub cmdBOk Click()
'retrieve the proportion from a dialogbox
Dim proportion As Double
proportion = Val(tbProportion.Text)
'checking the range of the value (< 0.5 max.)
If (proportion < 0\#) Or (proportion >= 0.5) Then
    proportion = 0.01
End If
'copying the full training set to the first sheet
Sheets("credit-german-train-full").Select
Range("A2:U301").Select
Selection.Copy
Sheets("credit-german-md").Select
Range("A2").Select
ActiveSheet.Paste
'number of cells to clear
Dim nbMd As Long
nbMd = Int(6000# * proportion)
'counter for the number of missing values
Dim counter As Long
counter = 1
'coordinates
Dim i As Long, j As Long
'initialize the random number generator
Randomize (100)
Do While (counter <= nbMd)
    'row
    i = 2 + Int(300 \# * Rnd)
```

```
'column
    j = 1 + Int(20 \# * Rnd)
    'checking if the cell is already empty
   If (Cells(i, j).Value <> "?") Then
        Cells(i, j).Value = "?"
        counter = counter + 1
   End If
Loop
'set the text file name
Dim nomFichier As String
nomFichier = "\credit-german-md.txt"
nomFichier = ThisWorkbook.Path + nomFichier
'checking if the file already exists
Dim fs As Variant
Set fs = CreateObject("Scripting.FileSystemObject")
If (fs.FileExists(nomFichier) = True) Then
   fs.deletefile (nomFichier)
End If
'save the sheet into a text file
ThisWorkbook.SaveAs fileName:=nomFichier, FileFormat:=xlTextWindows
'close the form (dialogbox for the proportion setting)
formMD.Hide
End Sub
```

We note that if we insert 300 missing value in our dataset, it does not mean that we have 300 observations with at least one missing value. Some instances have more than on missing value (e.g the row #13 into the screenshot above), others are complete. In the follows, we count the number of instances with no missing values in each generated learning samples.

### 3 Processing the missing values with R

#### 3.1 The R program

We wish to compare two methods of dealing with missing data: the listwise deletion and univariate imputation (replacement by the average for the quantitative variables, by the mode for the categorical ones). To evaluate the efficiency of the strategies, we apply the models learned from the pre-treated dataset on the same test sample with no missing values. The best strategy is the one that induces the best model i.e. which has the best test accuracy rate.

The R program is subdivided in several parts.

**Loading the samples**. We load the learning sample (with some missing values) and the test sample (with no missing values).

```
#loading the test sample
setwd("... votre répertoire ...")
data.test <- read.table(file="credit-german-test.txt",dec=".",sep="\t",header=T)
#loading the test sample</pre>
```

data.train <- read.table(file="credit-german-md.txt",dec=".",sep="\t",header=T,na.strings="?") summary(data.train)

**Listwise deletion approach**. The first strategy is to remove from the dataset the observations which contain at least one missing value. This is the listwise deletion strategy. We can reduce dramatically the dataset size with this approach. But in the MCAR context, with a moderate proportion of missing values, it can be acceptable.

#first approach: listwise deletion
data.train.omitted <- na.omit(data.train)
summary(data.train.omitted)
print(paste('Remaining observations : ',as.character(nrow(data.train.omitted)))))
model.omitted <- glm(classe ~ ., family = binomial, data = data.train.omitted)</pre>

**na.omit(.)** removes from the dataset the instances with at least one missing value. **model.omitted** is the model learned from the resulting observations.

**Univariate imputation**. We use the following program for the univariate imputation. We obtain the **model.imputed** model from the pre-treated learning sample.

```
#second approach: univariate imputation univariée (mean, mode)
data.train.imputed <- as.data.frame(lapply(data.train,traiter_missing_data))
summary(data.train.imputed)
model.imputed <- glm(classe ~ ., family = binomial, data = data.train.imputed)</pre>
```

The **traiter\_missing\_data(x)** function makes the distinction between the continuous (numeric) and the factor (categorical) variables.

```
#according to the variable type
traiter_missing_data <- function(x){
    if (is.factor(x) == T){
        return(traiter.discrete(x))
    } else {
        return(traiter.numeric(x))
    }
}</pre>
```

Thus, for the continuous attribute, we calculate the mean on the available values and we replace the NA by this value:

```
#continuous variable, replace NA by the mean
traiter.numeric <- function(x){
    y <- x
    z <- na.omit(y)
    if (length(z) < length(y)){
    m <- mean(z)
    y[is.na(y)] <- m</pre>
```

} return(y)

The approach is similar for the categorical variables; but we use the 'mode'<sup>6</sup> instead the mean:

```
#categorical variable, replace NA by the mode
traiter.discrete <- function(x){
    y <- x
    z <- na.omit(y)
    if (length(z) < length(y)){
      frequence <- table(z)
      m <- which.max(frequence)
      y[is.na(y)] <- levels(x)[m]
    }
    return(y)
}</pre>
```

**Model evaluation**. Last step of our experiment, we evaluate the models. We use a specific function for this. It takes as input the test sample and the model. It computes the model prediction, the confusion matrix and the accuracy rate. We use the accuracy rate instead of the error rate to obtain results directly comparable to the outputs of the other tools.

### #new.dataset is the test sample #model is the model to evaluate

```
pred_and_confusion_matrix <- function(new.dataset,model){
    #score predicted by the model</pre>
```

```
data.pred.prob <- predict(model,newdata = new.dataset)
```

#classification from the score

```
data.pred.class <- ifelse(data.pred.prob > 0.5,"B","A")
```

data.pred.class <- as.factor(data.pred.class)</pre>

```
#confusion matrix (observed class values vs. predicted value)
```

```
mc <- table(new.dataset$classe,data.pred.class)</pre>
```

```
print(mc)
#accuracy rate
```

```
ca <- (mc[1,1]+mc[2,2])/sum(mc)
```

```
print(ca)
```

### 3.2 Organization of the experiments

The idea is to observe the impact of the treatment strategy when the proportion of missing values increases. We try the following percentages: 0.5%, 1%, 2%, 5%, 10% and 20%.

<sup>&</sup>lt;sup>6</sup> http://en.wikipedia.org/wiki/Mode\_%28statistics%29

The reference result is the performance of the model learned from the complete training sample (0% missing data). Of course, both strategies should provide the same result since no missing value processing is performed. The test accuracy rate of the model is 73%. We should not obtain a better result for the model learned on the learning samples with missing values below (normally).



#### **3.3 Experiment results**

We obtain the results into the table below.

GERMAN D	ATASET	Acci	uracy rate
% missing	# complete obs.	Listwise Del.	Univ. Imputation
0,00%	300	0,7300	0,7300
0,50%	272	0,7100	0,7257
1,00%	246	0,7129	0,7286
2,00%	201	0,7214	0,7186
5,00%	111	0,6729	0,7114
10,00%	40	ERR	0,7086
20,00%	4	ERR	0,7214

We observe that:

- The presence of the missing values deteriorates the model performance.
- When there is a little proportion of missing values (up to 2%, about 201 complete instances), the two approaches give similar performances.

- When the proportion is higher, the listwise deletion approach seems inappropriate. Starting from a certain proportion (10%), the learning process is impossible because there are too few complete instances into the dataset.
- In contrast, the univariate imputation has a really good behavior. In the configuration MCAR, the strategy is appropriate even when there is 20% missing values in the training sample. In fact, this is not surprising. The "?" are well spread throughout the columns, there is finally a few proportion of missing observations for each variable. The value used for the replacement is viable. For the proportion of 20%, about 240 values are available in each column.

Again, we are in the perfect MCAR context in this tutorial. We must not forget it when we read the results. However, they are really interesting. Particularly, the univariate imputation seems viable in this context, even when the proportion of missing values is high.

### 4 Processing missing values in other software

In this section, we describe the tools for processing missing values in various data mining software. We work on the learning sample with 5% of missing values (111 complete instances on 300). The test sample is the same as previously.

#### 4.1 ORANGE

#### 4.1.1 Programming the analysis

**Missing value imputation**. After we launch Orange, we can define a new schema.



We set the FILE (DATA tab) component into the canvas. We set the file name (CREDIT-GERMAN-MD-5 00.TXT) and we specify the code '?' (DON'T KNOW) which represents the missing value.

We visualize the dataset with the DATA TABLE (DATA tab) tool.

Orange Canvas						
File View Options Widget Help	💾 Data Table					• X
	Info	cre	dit-german-md-5_00	(Examples)		
Data Visualize Classify Regression	300 examples, 189 (63.0%) with missing values		checking_status	duration	credit_history	рі 🔺
🕒 🔉 🖪 🛄 🔚	105 (03.0 %) with hissing values.	1	<0	12	critical/other ex	new ca
	20 attributes,	2	<0	6	? 🔶	furnitu
Data Table	no meta attributes.	3	<0	6	critical/other ex	new ca
•	Discrete class with 2 values.	4	no checking	15	existing paid	used c
		5	>=200	18	all paid	radio/1
	Settings           Image: Setting	6	no checking	18	all paid	new ca
		7	no checking	48	critical/other ex	busine
	Show attribute labels (if any)	8	0<=X<200	30	critical/other ex	new ca
	Resize columns: + -	9	no checking	18	delayed previo	busine
	Restore Order of Examples	10	no checking	18	critical/other ex	used c
	Colors	11	no checking	? 🔶	all paid	used c
Examples Data Table	Visualize continuous values	12	>=200	10	? 🔶	radio/1
	Color by class value	13	no checking	12	critical/other ex	new ca
	Set colors	14	no checking	24	existing paid	new ca
File		15	0<=X<200	13	existing paid	radio/1
	Selection	16	<0	18	critical/other ex	new ca
	Send selections	17	no checking	18	critical/other ex	radio/1
	Commit on any change	18	no checking	9	existing paid	radio/1
		19	<0	18	existing paid	new ca 🔻
	Report	•	III			F.

Orange announced that 189 observations contain at least one missing value (189 + 111 complete instances = 300). It handles properly the missing value code during the importation.

We insert the IMPUTE (DATA tab) tool into the canvas. We click on the OPEN menu to set the parameters. The following options are available:

- DON'T IMPUTE. Nothing is done. The subsequent learning algorithm (e.g. LOGISTIC REGRESSION) handles the missing values.
- AVERAGE /MOST FREQUENT. This is the univariate imputation described above.
- MODEL-BASED IMPUTER. It uses a nearest neighbor algorithm to replace the value of a variable from the values of the other variables. This approach seems better than the univariate one because we make use of the relation between the variables. But the calculation time may be prohibitive on a large dataset. This approach is known also as the hot deck imputation strategy.
- RANDOM VALUES. Orange tries to approximate the distribution of each variable. It then uses a value retrieved in accordance with approximated distribution into the range of possible values.
- REMOVE EXAMPLES WITH MISSING VALUES. This is the "listwise deletion" strategy.

We choose the MODEL BASED IMPUTER approach.

*N.B.:* We note that we can define a specific strategy for each variable of the database. We can even directly specify a value to replace the missing ones for each variable.

We add the LOGISTIC REGRESSION tool (CLASSIFY tab). We do not use the variable selection process.

Orange Canvas	. 0 %
File View Options Widget Help	
	D
Data Visualize Classify Regression Evaluate Un	
Learner/Classifier Name	
Bayes Regression Neighbours Tree Tr	e Viewer Tre
Stepwise attribute selection	4
Add threshold [%] 10 🌲	^
Remove threshold [%] 10 🔶	
Examples	
Data Table	
Examples Classification/Regression trees	=
File Examples	
Impute Logistic Regression Apply	
Report	
	<b>.</b>

We note that the component "logistic regression" has an internal mechanism for the imputation of missing values (IMPUTATION OF UNKNOWN VALUES). Various solutions are proposed: univariate imputation (average, minimum, maximum); multivariate imputation where it predicts missing values of a predictive variable from the other variables in the database (using a decision/regression tree). This feature is enabled if we choose the DON'T IMPUTE option into the IMPUTE component.

🤓 Orange Canvas	
<u>File View Options Widget Help</u>	File (2)
	Data File
Data Visualize Classify Regression	credit-german-test.txt 🔹 🚺 🐼 Reload
?	Info ?
File Paint Info Save Data	700 example(s), 20 attribute(s), 0 meta attribute(s). tize Prepre
	Classification; Discrete class with 2 val0e(s).
	Advanced settings
	Missing Value Symbols
Examples Data Table	Don't care:
Examples	Don't know:
File Examples	New Attributes
Impute Log	Create a new attribute when existing attribute(s)
	Have mismatching order of values     Jarve no common values with the new (recommended)
	Miss some values of the new attribute
	… Always create a new attribute
File (2)	Durit
	Keport

Model evaluation on the test sample. The model is automatically learned when we close the setting dialog (we can visualize it using the NOMOGRAM  $tool^7$ ). We want to evaluate its performance on the test set. We insert again the FILE component and we select the CREDIT-GERMAN-TEST.TXT data file.

Then we add the TEST LEARNER component (EVALUATE tab). We set

the following connections.



<sup>&</sup>lt;sup>7</sup> E.g. <u>http://data-mining-tutorials.blogspot.fr/2010/07/naive-bayes-classifier-for-discrete.html</u>

For the connection between the sample **FILE (2)** and the **TEST LEARNERS** component, we must specify that it concerns a separate test set (SEPARATE TEST DATA).

We open the TEST LEARNERS component. We specify that we evaluate the classifier on the test sample. The test accuracy rate is 72.43%.

TestLearners	
Sampling	Evaluation Results
Cross-validation	Method CA
Number of folds: 5	1 Logistic regression 0.7243 <
Ceave-one-out	
Random sampling	
Repeat train/test: 10 🚔	
Relative training set size:	
Test on train data	
Test on test data	
Apply on any change	
Apply	
Performance scores	
Classification accuracy	
Sensitivity =	
Area under ROC curve	
Information score	
Target class	
good 💌	
Report	

#### 4.1.2 Comparison of the imputation techniques

We want to compare the various imputation methods on our dataset. This is easy under Orange.

	Minpute	TestLearners
Image: Construct of the system of the sys	Default imputation method Don't Impute Average/Most frequent Model-based imputer Random values Remove examples with missing values Individual attribute settings Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status Checking.status	Evaluation Results       Cross-validation       Number of folds:       Leave-one-out       Random sampling       Repeat train/test:       Relative training set size:       57%
Lata idoe Attributes	Don't impute D credit, history D purpose G credit, amount D saving_status D employment D installment_commitment D presidence_since D roperty_magnitude G age D other_payment_plans Class Imputation ☐ Impute class values Send data and imputer Apply ✓ Send automatically Report	<ul> <li>Test on train data</li> <li>Test on test data</li> <li>Apply</li> <li>Apply</li> <li>Performance scores</li> <li>Classification accuracy</li> <li>Sensitivity</li> <li>Secificity</li> <li>Area under ROC curve</li> <li>Information score</li> <li>F-measure</li> <li>Imaget dass</li> <li>good</li> <li>Report</li> </ul>

Orange proposes a very practical feature. We leave open the TEST LEARNERS window and we interactively modify the imputation option into IMPUTE component. The results are refreshed when we modify the option (SEND AUTOMATICALLY must be selected). We obtain the following table:

Approach	Class. Accuracy
Don't impute (Internal method of Logistic Reg.)	0.7357
Average / Most Frequent	0.7271
Model-based imputer	0.7243
Random Values	0.7357
Remove Examples with missing values (listwise deletion)	0.6614

Apart from listwise deletion approach, all methods are similar on our dataset. Curiously, the accuracy rate in some situations is higher than the accuracy of the learned model on the complete dataset under R (73%). Perhaps R and Orange<sup>8</sup> logistic regression algorithms have not exactly the same characteristics. It is possible also that it was simply the consequence of fluctuations sampling. Anyway, we find that the listwise deletion approach is not appropriate when the proportion of missing values increases (5% for the experiments under Orange).

#### 4.2 KNIME

KNIME proposes the MISSING VALUES component for the treatment of missing data. In this section, we reproduce inthis section the analysis schema that we have defined under Orange.

**Listwise deletion strategy**. We import the CREDIT-GERMAN-MD-5\_00.TXT data file using the FILE READER tool (IO / READ). The missing value is symbolized by the character '?'.

									6	- • ×
File Edit View Search Run Node Help		r								
📑 🕶 🔚 💼 🛷 🕶 🖢 🖛 🏷	Q 💛 🛛	100% 🗸   🕶	🛕 Dialog - :	2:1 - File Reade	r					
🖄 Workflow Projects	🛆 0: Tree Again	🔺 *2: Missing	File	/	$\frown$					
	,		California							
		7	Setungs	Tow Variables	Memory Policy	1				
Missing data imputation	rile Keauer		Enter ASC	II data file locati	on: (press 'Ente	er' to update preview)				
I ree Again		Configure	vali	d URL: 3/md_e	xperiments_wit	h_orange/german/crea	dit-german-md-	5_00.txt 👻	Browse	
	II 🛕 🗳	Execute								
		Execute and Ope	Basic Sett	ings		Dracarva ucar catt	tinge for new lo	cation		
	Node1	Cancel	read r	ow IDs		Column delimiter:	ah> ▼		Advanced	
		Reset					d kelke		- Internetari	
		Edit Node Name	read o	olumin neadérs		gnore spaces and	u tabs			
▼		New Weekflew A				Java-style comme	ents	Single line	comment:	
IO 📩										
Read		Collapse into Me								
File Reader		Expand Meta No	Preview							
ARFF Reader		Show Flow Varia		Clic	k column heade	er to change column pr	roperties (* = n	ame/type user s	settings)	
CSV Reader			Row ID	S checkin	I duration	S credit history	S purpose	F credit	S savings	S employ.
Eine Keader	4	Cut	Row0	<0	12	critical/other exist	new car	3499	<100	1<=X<4
Duble Reader		Сору	Row1	<0	6	?	furniture/eq	1872	<100	? =
PHIL PMML Reader		Paste	Row2	<0	6	critical/other exist	new car	1361	<100	<1
Wite		Undo	Row3	no checking	15	existing paid	used car	3029	<100	4<=X<7
The Write			Row4	>=200	18	all paid	radio/tv	1445	no known sa	4<=X </td
Cache	KINIME Consoli	> Kedo	Row6	no checking	48	critical/other exist	husiness	7629	< 100	>=7
	Noloc X	Delete	Row7	0<=X<200	30	critical/other exist	new car	4249	<100	unemployed
Data Manipulation	*** Cor Ø	0 File Table	Row8	no checking	18	delayed previously	business	2169	<100	1<=X<4
O Data Views	********		Row9	no checking	18	critical/other exist	used car	?	no known sa	unemployed 👻
Statistics	Log file is	located at: C		•						P.
Mining	WARN Fil	e Reader								
Ke Meta										
🔁 Flow Control										
Ph Mire	•						0	K	Apply	Cancel
								··	CAAC	cuncer

<sup>8</sup> http://orange.biolab.si/doc/reference/Orange.classification.logreg/

We launch the importation b	clicking on the EXECUTE menu.
-----------------------------	-------------------------------

File Reader	Interactive Ta	ble					
<b>B</b>	<b>1</b>						
		A Table View	2.2 - Interactive	Table			- 0 X
Node 1	Node 2		v - 2.2 - Interactive	Table		6	
		File Hilite N	lavigation View C	Dutput			
		Row ID	S checkin	duration	S credit_history	S purpose	↓ credit
		Row0	<0	12	critical/other exist	new car	3499
		Row1	<0	6	?	furniture/eq	1872
		Row2	<0	6	critical/other exist	new car	1361
		Row3	no checking	15	existing paid	used car	3029
		Row4	>=200	18	all paid	radio/tv	1445
		Row5	no checking	18	all paid	new car	6458
		Row6	no checking	48	critical/other exist	business	7629
		Row7	0<=X<200	30	critical/other exist	new car	4249
		Row8	no checking	18	delayed previously	business	2169
		Row9	no checking	18	critical/other exist	used car	?
		Row 10	no checking	?	all paid	used car	7485
		Row11	>=200	10	?	radio/tv	1347
		Row 12	no checking	12	critical/other exist	new car	926
		Row13	no checking	24	existing paid	new car	1249
		Row 14	0<=X<200	13	existing paid	radio/tv	2101
		Row 15	<0	18	critical/other exist	new car	5302
		Row 16	no checking	18	critical/other exist	radio/tv	2238
		Row17	no checking	9	existing paid	radio/tv	2697
		Row18	<0	18	existing paid	new car	2249
		Row 19	no checking	12	existing paid	radio/tv	1262
		Row20	no checkina	24	all paid	business	1559
			∢ III		an para	Doom 1000	1005

To check the data importation, we visualize the dataset using the INTERACTIVE TABLE (DATA VIEWS) component

	🔺 Dialog - 2:3 - Missing Value
	File
	Default Individual Flow Variables Memory Policy
	Integer Columns
	Tataa
	i integer
	Do Nothing   Remove Row
	O Min O Max
	Mean Most Frequent
	Fix Value: 0
Interactive Table	Double Columns
	D Double
	O Nothing  Remove Row
File Reader / Missing Value	🔘 Min 💿 Max
	○ Mean
	Fix Value:
	String Columns
Node 1 Node 3	S String
	O Nothing  Remove Row
	Most Frequent
	Fix Value:
	Unknown Columns
	2 Unknown
	Do Nothing Remove Row
	OK Apply Cancel

We insert now the component MISSING VALUE (DATA MANIPULATION / COLUMN / TRANSFORM) into the workflow. We implement the listwise deletion strategy (REMOVE ROW), whatever the variable type, into the settings dialog (CONFIGURE menu). Again, to check the result of the treatment, we visualize the dataset with the interactive table component. We observe that the rows containing at least one missing value (ROW0, ROW1, ROW2,

#### ROW9, ROW10, etc.) are removed from the learning set.



**Naïve bayes classifier**. In the first version - in French - of this tutorial, I did not see that the Logistic Regression is into the STATISTIC branch and not into the MINING branch into the node repository. I have thought that this approach was not available. I decided to use the naive bayes classifier, which is also a linear classifier. Subsequently, Loïc Lucel (thank you very much Loïc) had pointed out to me the presence of logistic regression. So I decided to keep in this tutorial the studies for the naive bayes classifier and the logistic regression.

We add the NAÏVE BAYES LEARNER (MINING / BAYES) into the workflow. We set CLASSE as classification column.



We click on the EXECUTE AND OPEN VIEWS menu to obtain the results. Knime provides the conditional mean and standard deviation for continuous descriptors, the contingency table for the calculation of the conditional probabilities for the discrete ones.

A Naive Bayes Learner View - 0:5 - Naive Bayes Learner									
File									
Class counts for classe									
Class:		bad good							
Count:		35		76					
Total count: 111	Total count: 111								
Gaussian distribution for age per o	class value	Continu	ous de	scriptor					
		bad		good					
Count:		35		76					
Mean:		33.42857		38.07895					
Std. Deviation:		11.53584		11.65248					
Rate:		35/111		76/111					
P(checking_status   class=?)	Discret	e desci	riptor						
Class/checking_status	0=X200	0	>=200	no checking					
bad	12	12 12 3		8					
good	23	23 9 5 39							
Rate:	32%	32% 19% 7% 42%							
•									

**Evaluation on the test set**. We want to evaluate the model on the test set. We load this last one by using another FILE READER component. Then, we compute the predicted values using the NAIVE BAYES PREDICTOR tool.



To obtain the confusion matrix, we use the SCORER component. We set (CONFIGURE menu) CLASSE in row and the predicted values [WINNER (NAÏVE BAYES)] in column.



We click on EXECUTE AND OPEN VIEWS. The accuracy rate is 70.286%.



The naive bayes classifier is less penalized by deleting rows than logistic regression. Indeed, the accuracy rate of the model learned from the complete data (without missing values) is 72%. This behavior is probably the consequence of the low number of parameters to estimate from data.

*Note*: Of course, when we have a high proportion of missing values (e.g. 10%, it remains 40 rows into the learning set), even if the learning process is possible, the model is heavily deteriorated (accuracy rate: 63.429%).

Univariate imputation. When we choose the univariate imputation in the IMPUTE component (mean for integer and double columns, most frequent for the string ones), the accuracy rate becomes 71.571%. For the naive bayes classifier also, as the logistic regression, the univariate imputation is better than the listwise deletion strategy in the MCAR missing values context.

🛆 Confusion Matrix - 0:8 - Sco 🗖 🔲 💌									
File Hilite									
classe \ Wi	bad	good							
bad	103	107							
good	92	398							
Correct classified: 501 Wrong classified: 199 Accuracy: 71.571 % Error: 28.429 %									





The categorical variables are automatically coded. The output complies to the standards of the domain. We have the estimated parameters, their standard error and test of significance statistic.

🔔 Logistic Regression Result View - 0:9 - Logistic Regression (Learner)									
File									
Statistics on Logistic Regression									
Logit	Variable 🧲	Coeff.	Std. Err.	z-score	<b>P</b> > z				
bad	checking_status=0	1.2391	0.4972	2.4919	0.0127				
	checking_status=>=200	-0.3502	0.9284	-0.3772	0.706				
	checking_status=no checking	-1.3266	0.5226	-2.5383	0.0111				
	duration	0.006	0.0192	0.3108	0.756				
	credit_history=critical/other existing	-2.8089	0.9262	-3.0328	0.0024				
	credit_history=delayed previously	-2.415	1.008	-2.3959	0.0166				
	credit_history=existing paid	-2.6995	0.8692	-3.1059	0.0019				
	credit_history=no credits/all paid	-0.1139	1.3018	-0.0875	0.9303				
	purpose=domestic appliance	0.8377	1.6268	0.5149	0.6066				
	purpose=education	1.6245	1.0658	1.5241	0.1275				
	purpose=furniture/equipment	-0.3342	0.8601	-0.3886	0.6976				
	purpose=new car	0.7304	0.7587	0.9627	0.3357				
	murposo-other	0 7004	1 612	0 4242	0.6641	-			

We have launched the workflow on the learning set containing 5% of missing values. The test error rate is **72.286%** when we use the univariate imputation.



With the listwise deletion strategy, it becomes **67.286%**.



#### 4.3 RapidMiner

Two tools are needed for the detection of missing values in RapidMiner. Thereafter, we use a specific component for their treatment. RapidMiner proposes the listwise deletion and the univariate imputation strategies.

**Detection of missing values**. After starting RapidMiner, we create a new "Process". We use the READ CSV (IMPORT/DATA) component to import the dataset. The easiest way is to use the wizard (IMPORTATION CONFIGURATION WIZARD button) to set the parameters of the importation.

🚯 Missing values* – RapidMiner@Maison-PC		1	
<u>File Edit Process Tools View H</u> elp			
🖹 📦 🔒 🕞 🗊 🔊 🔌	🕨 📗 🛐 🛒 🖲		
Overview 🛛 💱 🖨 🖻	Process X EXML X		B Parameters × 52 🗢 🔟
Repositories 💥	Vicess V		🍓 😼 🦻 😾 🦌 🖌
Ø ← [Filter] Ø ▶ ↓		8	🎢 🎢 Import Configuration Wizard
	Read CSV		csv file
⊡ 🔄 Import (26) ⊡ 🔄 Data (18)	fil 🦂 out		column separators
- A Read CSV - Read ARFF			Trim linas
- Read Excel with Format			Comment      Comment
Read Access     Read AML     Read XML     Read XRFF     Read Database			Read CSV
- A Stream Database - A Read SPSS	A Problems X Colon X		Synopsis This operator can read cay files
– 🥔 Read Stata – 🛃 Read Sparse	No problems found		
- 🏕 Read DBase	Message	Fixes Location	Description This operator can read csy files, where all values of an example are
– 🦂 Read BibTeX			writen into one line and separated by an constant separator. The
Read URL			separator might be specified in the column separators parameter. The default will split the line on each comma, semicolon and blank.
Models (2)			Arbitrary regular expressions are usable as separator. Empty values
Results (1)			and the question mark will be read as Missing Values. You can
0			

Defining the type of the variable is essential. For our dataset, all the continuous attributes are REAL; the others are POLYNOMIAL (more than two categories) or BINOMIAL (two categories). In addition, we must specify the label column (CLASSE).

🚯 Data	import wizard	d - Step 4 of 4										<b></b> >	x
Ž	This wizard guides you to import your data. Step 4: RapidMiner uses strongly typed attributes. In this step, you can define the data types of your attributes. Furthermore, RapidMiner assigns roles to the attributes, defining what they can be used for by the individual operators. These roles can be also defined here. Finally, you can rename attributes or deselect them entirely.												
	<u>R</u> eload data	<u>Gues</u>	ss value types	Previe	w uses only fi	rst 100 rows.	<u>D</u> ate format			•			
	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$		-
nal_st	other_partie	residence_s	property_ma	age	other_paym	housing	existing_cre	job	num_depen	own_telepha	foreign_worl	classe	
0 🔻	polyno 🔻	real 💌	polyno 🔻	real 💌	polyno 🔻	polyno 🔻	real 💌	polyno 🔻	real 💌	polyno 🔻	polyno 🔻	binomi 🔻	ŀ
ute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	attribute 🔻	label 🔻	1
ə div/dı	co applicant	2	real estate	29	none	?	2	skilled	1	none	yes	bad	
single	none	4	no known pr	36	none	for free	3	high qualif/s	1	yes	yes	good	
												$\rightarrow$	
📀 0 ei	rors.									🖌 Ignore	e errors 📃 S	how only <u>e</u> rror	rs
	Row,	Column			Error			Original value			Message		
	Previous Next Einish Scancel												

RapidMiner uses this first step to detect missing values. For the REAL columns, the character "?" is inconsistent with the definition of the variable. The software deducts that there are missing values. We observe this by running the PROCESS (PROCESS / RUN menu).

In the EXAMPLE SET (READ CSV) results tab, it lists the number of missing observations for each REAL variable (e.g. 17 for DURATION, 20 for CREDIT\_AMOUNT, etc.).

🥸 Missing values – Rap	pidMiner@Maison-PC			100			x			
Eile Edit Process Tools View Help										
🕒 😋 📰 🕼 🖉 🖍 🔌 🕨 💷 🛐 🛒 💿										
🔀 Result Overview 🕱 🖉 📳 ExampleSet (Read CSV) 🕱										
💿 Meta Data View 🔵 Data View 🔵 Plot View 🔵 Annotations 📘 🖬 🥔										
ExampleSet (300 exan	nples, 1 special attribute	, 20 regular attributes)			1	- 🏥 👻				
Role	Name	Туре	Statistics	Range	Missin	gs				
label	classe	binominal	mode = good (210), le	bad (90), good (210)	0					
regular	checking_status	polynominal	mode = no checking (1	<0 (85), no checking (1	0					
regular	duration	real	avg = 22.512 +/- 13.23	[4.000 ; 60.000]	17 🔶 🗕					
regular	credit_history	polynominal	mode = existing paid (	critical/other existing (	0					
regular	purpose	polynominal	mode = new car (84), I	new car (84), furniture/	0					
regular	credit_amount	real	avg = 3709.750 +/- 328	[276.000 ; 15857.000]	20					
regular	savings_status	polynominal	mode = <100 (179), le	<100 (179), no known	0					
regular	employment	polynominal	mode = 1<=X<4 (93), I	1<=X<4 (93), ? (12), <1	0					
regular	installment_commitm	real	avg = 3.045 +/- 1.098	[1.000 ; 4.000]	10 🔶 💳					
regular	personal_status	polynominal	mode = male single (1	female div/dep/mar (92	0					
🔒 Log 💥 💱 🖨										
🔒 🥔 🤻						nc	nitor			
Dec 3, 2011 5:49:19 AM CONFIG: Loading perspectives. Dec 3, 2011 5:49:20 AM INFO: Checking for updates. Dec 3, 2011 5:49:21 AM INFO: Checking for updates since Wed Nov 30 08:59:11 CET 2011. Dec 3, 2011 5:49:24 AM INFO: Decoupling process from location //NewLocalRepository/Missing values. Process is now associated with file //NewLocalRepository/Missing values. Dec 3, 2011 5:52:27 AM WARNING: Password in XML file looks like unencrypted plain text. Dec 3, 2011 5:53:02 AM INFO: Saved process definition at //NewLocalRepository/Missing values'. Dec 3, 2011 5:58:17 AM INFO: Reading example set										

We come back to the PROCESS window. We insert the DECLARE MISSING VALUES component into the workspace. We set that we want to treat only nominal variables (because the detection for the continuous attribute is already completed during the importation process) and we specify the code to missing values ("?").

😢 Missing values – RapidMiner@Maison-	PC				- • ×
<u>File E</u> dit <u>P</u> rocess <u>T</u> ools <u>V</u> iew <u>H</u>	elp				
📔 📦 📰 🔛 🔊 🔊	a 🕨 🛯 🔳 🛐 🖉 🛈				
🔎 Overview 🐹 🚰 🖨 🔟	🥜 Process 💥 👍 XML 💥			Parameters	X 53 0 D
	← → → → ↑ Process →	að 🗕 🎲 📗	I 🖾 🦂 🗸	🚨 🕤 🔜 🗩	🕵 🔖 🕶
Repositories 🔀	Main Drocoss		<u> </u>	💻 Declare N	lissing Value 🛛 🚽
🔊 Operators 🛛 🤣 🗣 🚛	Read CSV			attribute filter type	value_type 💌
□- 🏐 Data Transformation (3) □- 🧐 Value Modification (1)	e Till Sour			value type	nominal
- III Declare Missing Value				🔲 use value type	exception
Replace Missing Values	inp D	exa exa		invert selection	ı
		•		🔲 include specia	I attributes
				mode	nominal 🔻
				nominal value	?
	A Problems 🗶 👌 Log 🗶			-	
	No problems found				
	Message	Fixes	Loation		
				$\sim$	
0					

😵 Missing values – RapidMiner@Maison-PC							×		
Eile Edit Process Tools View Help									
🕒 📦 🔚 🖪	🕒 😋 📰 🖾 🔊 🔊 🔺 🕨 📗 🔟 🖳 💱 😨								
🦷 🔀 Result Overview	🗴 🗶 🗍 📑 ExampleSe	t (Declare Missing Value	e) 🗶 🔪						
💿 Meta Data View) 🔘	Data View 🔘 Plot Viev	v 🔘 Annotations			_	i.	٠ 🌭		
ExampleSet (300 exam	nples, 1 special attribute	, 20 regular attributes)					- 🗉		
Role	Name	Туре	Statistics	Range	- 🕂 Miss	ings			
label	classe	binominal	mode = good (210), le:	bad (90), good (210)	0				
regular	checking_status	polynominal	mode = no checking (1	<0 (85), no checking (1	14				
regular	duration	real	avg = 22.512 +/- 13.23	[4.000;60.000]	17				
regular	credit_history	polynominal	mode = existing paid (	critical/other existing ({	19				
regular	purpose	polynominal	mode = new car (84), I	new car (84), furniture/	18				
regular	credit_amount	real	avg = 3709.750 +/- 328	[276.000 ; 15857.000]	20				
regular	savings_status	polynominal	mode = <100 (179), le:	<100 (179), no known :	11				
regular	r employment polynominal mode = 1<=X<4 (93), h 1<=X<4 (93), ? (0), <1 ( 12								
regular	installment_commitme	real	avg = 3.045 +/- 1.098	[1.000 ; 4.000]	10		$\overline{}$		
🔒 Log 🗙 💱 🖨									
mitor									
Dec 3, 2011 6:11:38 AM INFO: Process //NewLocalRepository/Missing values starts									
Dec 3, 2011 6:11:38 AM INFO: Loading initial data.									
Dec 3, 2011 6:11:38 AM INFO: Saving results.									
Dec 3, 2011 6:11:38 AM INFO: Process //NewLocalRepository/Missing values finished successfully after 0 s									
Dec 3, 2011 6:31:43 AM INFO: Process // VewLocal Reposition / Maing Studies starts									
Dec 3, 2011 6:31:43 AM INFO: Loading initial data.									

We select the DATA VIEW option into the results tab. We can filter the dataset in different ways. We select for instance the rows with at least one missing value, we obtain 189 rows.

😵 Missing values – RapidMiner@Maison-PC											
<u>F</u> ile <u>E</u> dit	Process To	iols ⊻iew <u>H</u>	elp								
i 🗋 📹	Solution     Solution										
🛛 🛒 Resu	lt Overview 🛛	👔 Exam	pleSet (Decla	are Missing Va	lue) 🛛 🔪						
🔘 Meta Dat	a View 💿 Da	ita View 🔵 Pl	ot View 🔘 A	nnotations						i.	🐣 👻
ExampleSet	(300 example	s, 1 special at	tribute, 20 reg	ular attributes	)			iew Filter (189	9 / 300): miss	ing_attributes	ş 🔻
Row No.	classe	checking_st	duration	credit_history	purpose	credit_amo	savings_st	employment	installment	personal_st.	.other
1	bad	<0	12	critical/other	new car	3499	<100	1<=X<4	3	female div/d	co at
2	good	<0	6	? 🔶	furniture/equ	1872	<100	? 🔶	4	male single	none
3	good	<0	6	critical/other	new car	1361	<100	<1	2	male single	none
4	good	no checking	18	critical/other	used car	?	no known sa	unemployed	2	male single	none
5	bad	no checking	?	all paid	used car	7485	no known sa	unemployed	4	female div/d	none
6	good	>=200	10	?	radio/tv	1347	no known sa	4<=)<<7	4	?	none
7	good	0<=X<200	13	existing paid	radio/tv	2101	<100	<1	2	female div/d	guar
8	good	no checking	18	critical/other	radio/tv	2238	<100	1<=X<4	2	female div/d	none 🤜
- Log	x 53 o 🖻										
Monitor								onitor 🖹			
Dec 3, 2011 6:11:38 AM INFO: Process //NewLocalRepository/Missing values starts											
Dec 3, 2011 6:11:38 AM INFO: Loading initial data.											
Dec 3, 2011 6:11:38 AM INFO: Saving results.											
Dec 3, 2011 6.11.36 AW INFO: No Process // NReWLocal Repository/Missing Values initiated successituity after 0 s Dec 3, 2011 6.31.43 AW INFO: No filename of the for result file using stadult for longing results!											
Dec 3, 2011 6:31:43 AM INFO: Process //NewLocalRepository/Missing values starts											
Dec 3, 2011 6:31:43 AM INFO: Loading initial data.											

**Univariate imputation**. We add the REPLACE MISSING VALUES in our process (DATA TRANSFORMATION / DATA CLEANSING). RapidMiner proposes the AVERAGE for the missing values

<sup>&</sup>lt;sup>9</sup> The whole process seems complicated for a text file such as we want to handle in this tutorial. It becomes easier when we handle a dataset from a DBMS such as Oracle, etc. See <a href="http://www.youtube.com/watch?v=0IVZmAk0p14">http://www.youtube.com/watch?v=0IVZmAk0p14</a>

imputation. In reality, this option operates also for the categorical variables. In this situation, RapidMiner uses the MODE.

🚯 Missing values – RapidMiner@Maison-PC							
Elle Edit Process Iools View Help							
🖹 🔄 🖩 🕼 🔊 \land 🕨 📕 🚺 🖉 🕸							
C <sup>o</sup> Process 🕱 🚡 XML 🕱							
🗋 🖓 🕶 🗮 🔄 🖛 🗮 🔄 🖛 🗣 🔹 🏥 Process 🕨 🛛 🛷 🔹 🕼 👻	Parameters 🛛 🖧 🗇 🖻 🔪						
	🍒 😼 🖉 🦻 🦉						
Operators X Main Process	Replace Missing Values						
	create view						
E Value Modifica	attribute filter t						
exa exa exa							
	Invert selection						
	🔲 include special attributes						
	default						
	average -						
No problems found	columns 🛛 🖓 Edit List						
message Fixes LUCdUUI							

We want to insert the supervised learning component in our process. The logistic regression of RapidMiner does not correspond to the usual method that we find in statistical tools. It seems preferable to use the Naive Bayes classifier here.

Read CSV file out exa exa ori pre Naive Bayes tra mod exa

We launch the process. Into the tabledescribingtheconditional

distributions, we observe that the "?" value is not present. The imputations are really completed.

🚯 Missing va	lues – RapidN	liner@Maison	-PC	diam'r brant			
<u>F</u> ile <u>E</u> dit	<u>P</u> rocess <u>T</u> o	ols ⊻iew <u>H</u>	elp Nouve	le capture Ctrl+N			
🎴 📹		<b>9</b>	<b>A</b>	II 🔲 🕅 🛒 🛈			
🛛 🔀 Resul	t Overview  🛛	🍸 💡 Simp	leDistribution	Naive Bayes) 🚿 🗌			
O Text View	O Plot View	Distribut	on Table) 🔘 .	nnotations	🖺 🌛 🗸		
Attribute	Parameter	bad	good				
checking_st;	value=<0	0.489	0.195				
checking_st;	value=no chi	0.167	0.505				
checking_st;	value=>=200	0.033	0.062				
checking_st;	value=0<=X∘	0.311	0.238				
checking_st;	value=?	0.000	0.000	_			
checking_st;	value=unknc	0.000	0.000				
duration	mean	25.345	21.298				
duration	standard dev	13.375	12.462				
credit_histor	value=critica	0.178	0.329				
credit_histor	value=?	0.000	0.000 🧲	-			
credit_histor	value=existir	0.533	0.533				
nitor							
Dec 3, 2011 6:10:58 AM INFO: Loading initial data.							
Dec. 3. 2011 6:10:58 AM SEVERE: Process failed: operator cannot be executed (Extraction of nominal example value for							
•							

**Model evaluation**. We insert again the READ CSV component to load the test set (CREDIT-GERMAN-TEST.TXT). We make the following connections between the components.



Last, the PERFORMANCE CLASSIFICATION (EVALUATION / PERFORMANCE MEASUREMENT / CLASSIFICATION AND REGRESSION) component enables to compute the accuracy rate (ACCURACY).



We launch the process. The test accuracy rate is 74.86%.

🐼 Missing values – RapidMiner@Maison-PC									
Eile Edit Process Tools View Help									
P 👒 📰 🕞 🕬 🗪 🐴 🕨 🔳 🖏 🐺 👁									
🦷 🛒 Result Overview 🖇	🛛 🔏 PerformanceVector (Perform	mance) 🔀 🔪							
💿 Table / Plot View 🔵 T	Fext View 🔘 Annotations				🖬 🤳 👻				
Criterion Selector	Multiclass Classification Performance	mance 🔘 Annotations			🖪 🗛				
accuracy	Table View      Plot View								
	accuracy: 74.86%								
		true bad	true good	class precision					
	pred. bad	91	57	61.49%					
	pred. good 119 433 78.44%								
	class recall 43.33% 88.37%								
🕞 🥔 🤲									
Dec 3, 2011 7:03:38 AM INFO: Saving results.									
Dec 3, 2011 7.03.38 AM II	Cost 2011 1-00-00 minition - 1-00-00 minition - 1-00-00 ministricul Subcessiulity allel 0.5								

## 5 Results on the WAVE dataset (2 classes version)

To confirm the results highlighted in this tutorial, we repeat the same process on the binary version of the waveform<sup>10</sup> dataset (the instances corresponding to the third class are removed). We have 500 instances in the learning set and 32867 instances in the test set. The estimation of the accuracy rate will be more precise.

WA	VE DATASET	Accuracy rate			
% missing	# complete obs.	Listwise Del.	Univ. Imputation		
0,00%	500	0,9150	0,9150		
0,50%	451	0,9147	0,9150		
1,00%	405	0,9082	0,9162		
2,00%	331	0,9068	0,9160		
5,00%	177	0,8731	0,9174		
10,00%	72	0,7847	0,9192		
20,00%	6	ERR	0,9188		

We obtain the following results using the logistic regression:

The results are consistent with those obtained on the GERMAN dataset. When the proportion of missing values increases, the univariate imputation is far better than the listwise deletion strategy.

Curiously, compared with the performance of the model learned from the complete dataset, the accuracy rate of models seems improved when the missing values are replaced by the mean. I have no veritable explication for that. This is probably a consequence of sampling fluctuations. Anyway, the programs and the dataset are available. The reader can repeat the experiments.

<sup>&</sup>lt;sup>10</sup> http://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+%28Version+1%29

# 6 Conclusion

The treatment of missing data is a very difficult problem. We must make choices based on factors that we do not control very well (the missingness mechanism). In this tutorial, we have tried to show several software solutions. When the missing values are MCAR, univariate imputation (mean / mode) seems have a good behavior in the logistic regression context, to the extent that our main criterion is the generalization accuracy rate.

# 7 References

Allison, P.D. (2001), « Missing Data ». Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

Little, R.J.A., Rubin, D.B. (2002), « Statistical Analysis with Missing Data », 2<sup>nd</sup> Edition, New York: John Wiley.